TeleTrusT-Konferenz 2019

Bundesverband IT-Sicherheit e.V. (TeleTrusT)

Berlin, 28.11.2019

IT-Sicherheit & Maschinelles Lernen

Dr. Sven Herpig, Stiftung Neue Verantwortung

Maschinelles Lernen in Safety-kritischen Bereichen

Gesichtserkennung in Überwachungskameras
Krisenvorhersage/ -prevention
Judikative Prozesse (inkl. Übersetzungen)
Prozessoptimierung in kritischen Infrastrukturen

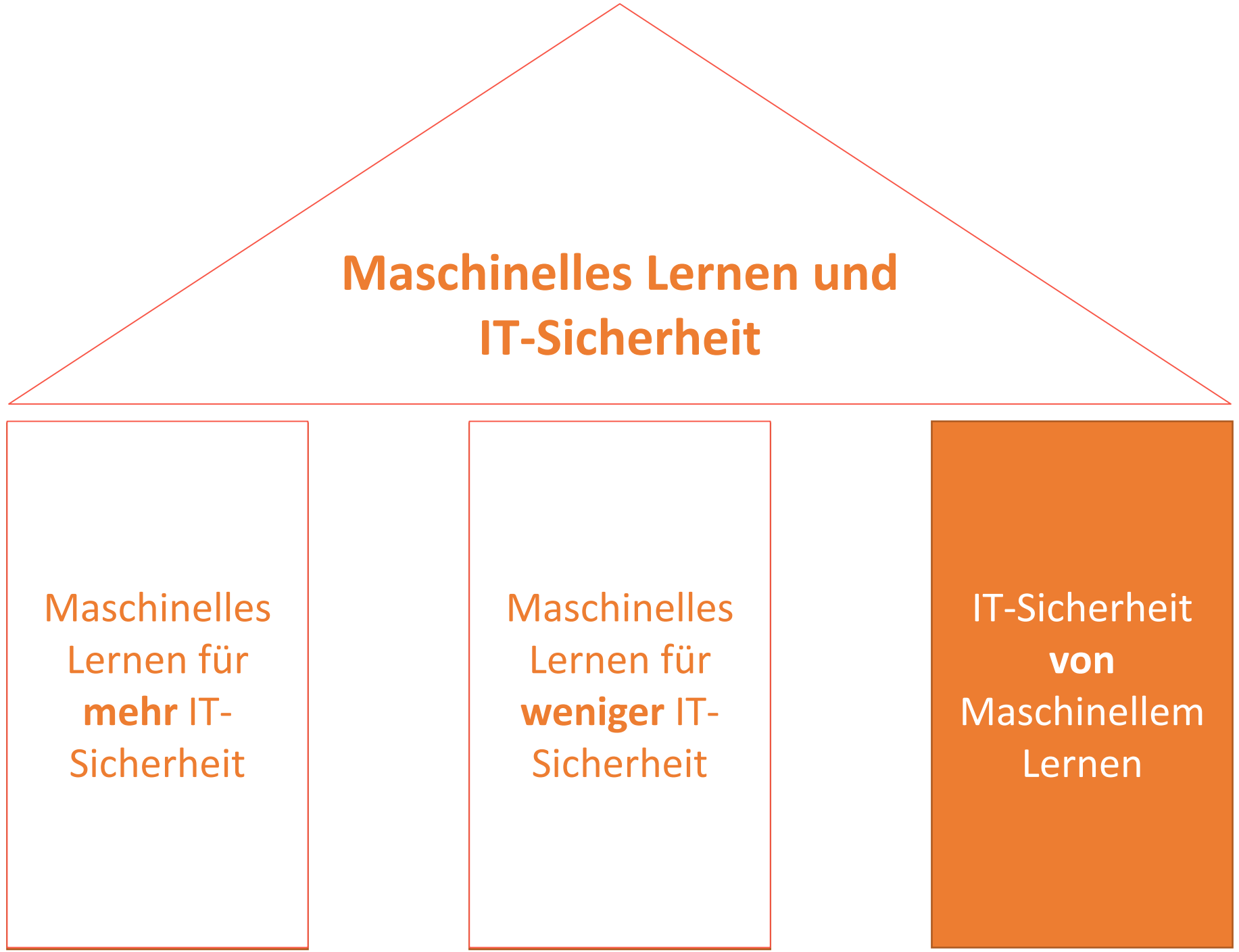
Nachrichtendienstliche Aufklärung, Datensammlung und Auswertung

Propaganda/ Desinformationskampagnen
Militärische Entscheidungsfindung und Logistik
Simulationen und Training

Steuerung von unbemannten militärischen Systemen (Fahrzeuge, Drohnen usw.)
Semi-autonome lethale Waffensysteme und Gegenmaßnahmen

Offensive und defensive Cyberoperationen
Gegenmaßnahmen zu ML-Systemen

u. v. m.


Maschinelles Lernen und IT-Sicherheit
Maschinelles Lernen für **mehr** IT-Sicherheit
Maschinelles Lernen für **weniger** IT-Sicherheit
IT-Sicherheit **von** Maschinellem Lernen

October 2019 · Dr. Sven Herpig

Securing Artificial Intelligence

Part 1: The attack surface of machine learning and its implications

Stiftung
Neue
Verantwortung

Think Tank für die Gesellschaft im technologischen Wandel

“Wenn Du nicht mehr weiter weißt, gründe einen Arbeitskreis!”

Transatlantic Cyber Forum [TCF] Working Group on “Information Security of Machine Learning” (gegründet 01/2019)

Charles-Pierre Astolfi, French Digital Council
Daniel Castro, Center for Data Innovation
Thomas “halvarflake” Dullien, optimyze.cloud
Joseph L. Hall, Center for Democracy & Technology
Maximilian Heinemeyer, Darktrace
Wyatt Hoffman, Carnegie Endowment for International Peace
Sven Jacob, BSI
Lydia Kostopoulos, ESMT Berlin
Johannes Otterbach, OpenAI
Thomas Reinhold, cyber-peace.org
Matthias Schulze, SWP
Sven Weizenegger, SUZA.io

Cathleen Berger, Mozilla
Betsy Cooper, Aspen Tech Policy Hub
Kenneth Geers, Atlantic Council [Fellow]
Mary Hanley, The University of Chicago
Ariel Herbert-Voss, Harvard University
Michael Hsieh, Transformative Cyber Innovation Lab
Frederike Kaltheuner, Privacy International
Igor Mikolic-Torreira, Georgetown University
Jörg Pohle, Humboldt Institute for Internet and Society
Kate Saslow, Stiftung Neue Verantwortung
Caroline Sinders, Harvard Kennedy School

28.11.2019, TeleTrusT-Konferenz 2019

Angriffsoberfläche von Maschinellem Lernen

Stiftung Neue Verantwortung

Think Tank für die Gesellschaft im technologischen Wandel

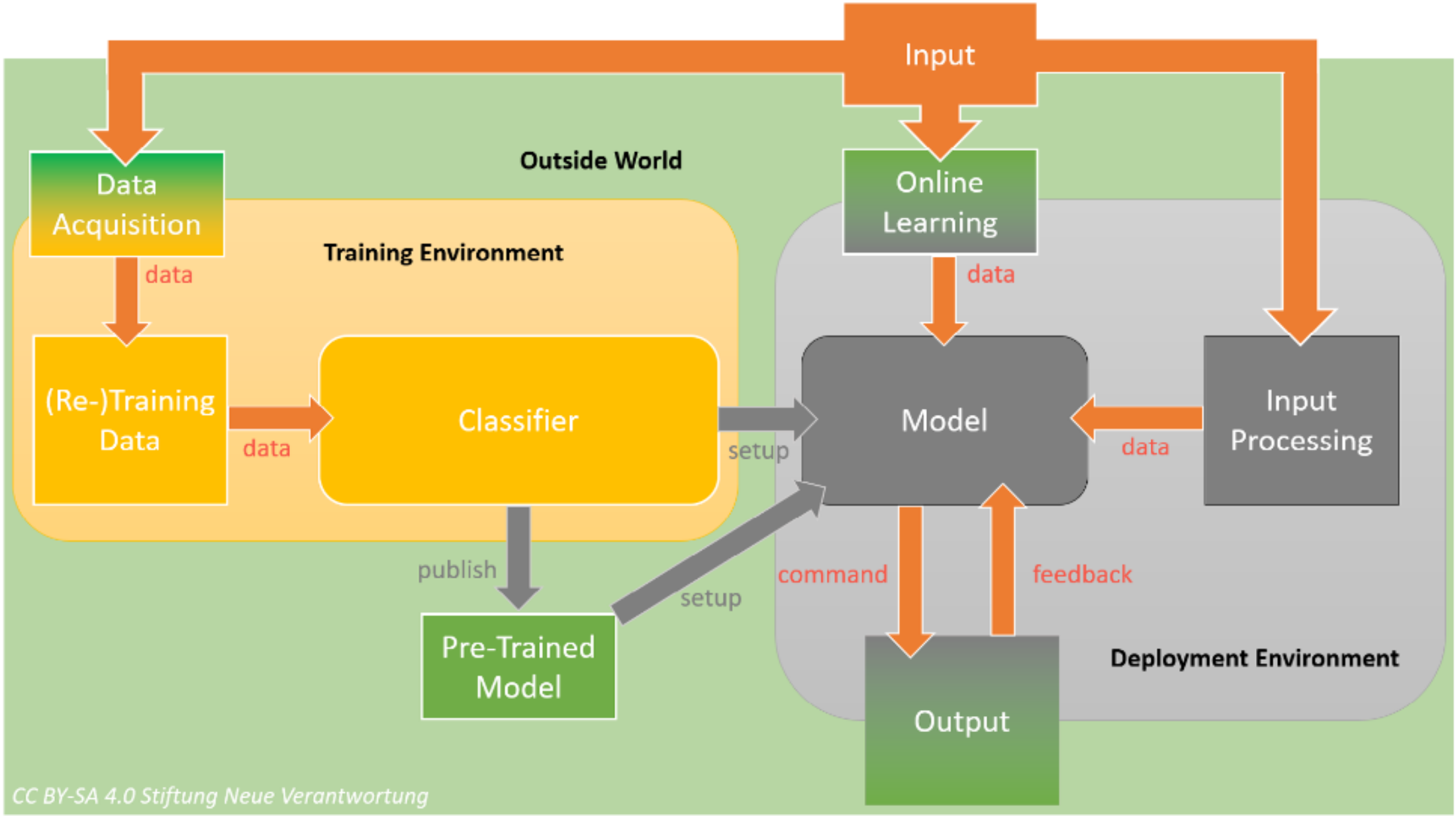


Figure 2: Attack surface of machine learning

Beispiel: ML-spezifische Supply Chain

1. Trainingsdaten per se (z. B. Verschlusssachen)
2. Datenbeschaffung (3rd Party oder 4th Party)
3. Datenverarbeitung (z. B. Outsourcing vom Labeling)
4. Trainieren des Models (z. B. unter Nutzung von Cloud-Anbietern)
5. Vortrainierte Modelle (u. a. von GitHub)
6. Live-Trainingsdaten (u. a. bei Online Learning/ Federated Learning)

Erkenntnisse – nicht ganz neu, aber teilweise anders

1. Die Angriffsoberfläche ist weitläufig und teilweise schwer bis gar nicht zu kontrollieren, was auch bei der Detektion und Attribution von Angriffen zu zusätzlichen Herausforderungen führt

2. Durch Angriffe auf ML-Systeme können verschiedenste Ziele erreicht werden, die möglicherweise -- auch wegen der Geschwindigkeit, Automatisierung und Skalierung -- zu Domino-Effekten führen können.

3. Das richtige “Threat Modeling” ist ein Kernstück um die IT-Sicherheit von ML-Systemen zu gewährleisten (u. a. ML-spezifische Supply Chain oder eingestufte Trainingsdaten)

4. Eine umfassende Betrachtung, sowohl vertikal (konventionelle IT-Sicherheitsherausforderungen) als auch horizontal (offensive und defensive Einsatzszenarien von ML), und eine klare Priorisierung der Schnittstelle von IT-Sicherheit und Maschinellen Lernen in der Nationalen KI-Strategie sind zwingend notwendig

“Even though it is difficult to predict whether information security will become a precondition for the successful development of machine learning going forward, securing machine learning, especially when it comes to [safety-critical deployment environments], is indispensable”

Securing Artificial Intelligence

Part 2: Policy Recommendations To Better Secure Machine Learning

Quo Vadis – “Forschung”

1. ML-spezifische Penetration Testing Fähigkeiten
2. Datenvalidierungsmethoden
3. Domänen-spezifische IT-Sicherheitsstandards
4. Eingebaute forensische Fähigkeiten
5. Robustheit von Algorithmen
6. Erklärbare/ Interpretierbare Künstliche Intelligenz
7. Human-in-the-loop und Human-on-the-loop
8. Redundante Algorithmen
9. Verantwortungsbewusster Umgang mit Schwachstellen bei ML
10. Supply Chain Absicherung

u. v. m.

Quo Vadis – “Politik”

Enquete-Kommission “Künstliche Intelligenz” des Deutschen Bundestags
- Bericht der Arbeitsgruppe “KI & Staat”

Deutsches Institut für Normung (DIN), Deutsche Kommission Elektrotechnik (DKE) und Bundesministerium für Wirtschaft und Energie (BMWi)
- Auftaktveranstaltung zur Arbeitsgruppe “IT-Sicherheit bei KI-Systemen” (10/2019)

European Telecommunications Standards Institute (ETSI)
- Industry Specification Group “Securing Artificial Intelligence” (09/2019)

Plattform Lernende Systeme
- Bestandsaufnahme und Lösungsansätze “Künstliche Intelligenz und IT-Sicherheit” (04/2019)

Stiftung Neue Verantwortung:
- Problemanalyse “Attack Surface of Machine Learning and Its Implications” (10/2019)
- Beitrag “IT-Sicherheit von Maschinellem Lernen” für die AG “KI & Staat” der Enquete-Kommission
- Agenda Setting auf deutscher und europäischer Ebene
- Teil 2 der “Securing Artificial Intelligence” Reihe, vermutlich in 10/2020

u. v. m.

As the day draws closer when the impact of human-made climate change will render the planet uninhabitable, it is perfectly understable to be less worried about cybersecurity.

Dr. Sven Herpig
Leiter “Internationale Cybersicherheitspolitik”
sherpig@stiftung-nv.de
@z_edian (Twitter)

Stiftung Neue Verantwortung

Think Tank für die Gesellschaft im technologischen Wandel